Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre

Citation for the original published paper (version of record):

Ahrens, J., Andersson, C. (2019). Perceptual evaluation of headphone auralization of rooms captured with spherical microphone
arrays with respect to spaciousness and timbre. Journal of the Acoustical Society of America, 145(4): 2783-2794. http://dx.doi.org/10.1121/1.5096164

Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre

Jens Ahrens, and Carl Andersson

Citation: The Journal of the Acoustical Society of America **145**, 2783 (2019); doi: 10.1121/1.5096164
View online: https://doi.org/10.1121/1.5096164
View Table of Contents: https://asa.scitation.org/toc/jas/145/4
Published by the Acoustical Society of America

Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre

Jens Ahrens[a] and Carl Andersson
Audio Technology Group, Division of Applied Acoustics, Chalmers University of Technology, 412 96 Gothenburg, Sweden

(Received 12 August 2018; revised 8 January 2019; accepted 8 January 2019; published online 30 April 2019)

A listening experiment is presented in which subjects rated the perceived differences in terms of spaciousness and timbre between a headphone-based headtracked dummy head auralization of a sound source in different rooms and a headphone-based headtracked auralization of a spherical microphone array recording of the same scenario. The underlying auralizations were based on measured impulse responses to assure equal conditions. Rigid-sphere arrays with different amounts of microphones ranging from 50 to up to 1202 were emulated through sequential measurements, and spherical harmonics orders of up to 12 were tested. The results show that the array auralizations are partially indistinguishable from the direct dummy head auralization at a spherical harmonics order of 8 or higher if the virtual sound source is located at a lateral position. No significant reduction of the perceived differences with increasing order is observed for frontal virtual sound sources. In this case, small differences with respect to both spaciousness and timbre persist. The evaluation of lowpass-filtered stimuli shows that the perceived differences occur exclusively at higher frequencies and can therefore be attributed to spatial aliasing. The room had only a minor effect on the results. https://doi.org/10.1121/1.5096164

[NX] Pages: 2783–2794

I. INTRODUCTION

Headphone renderings of spherical microphone array recordings constitute the audio equivalent of a panoramic video rendered on a head mounted display and can be a valuable tool in virtual and augmented reality applications. The underlying theory is well-known but practical implementations have only been available recently.

The term *rendering* refers to the auralization of a description of a sound scene. This scene description can be abstract or data-based. Rendering sometimes refers to the computation of loudspeaker signals, which can be presented by both ear-related loudspeakers, i.e., headphones, or by room-related loudspeakers (Blauert and Rabenstein, 2010). In the present context, one may also speak of *sound field synthesis* or *sound field re-synthesis*. Classical sound field synthesis aims at physically synthesizing a sound field over an extended area by means of arrays of loudspeakers (Ahrens, 2012). In the context of this article, the sound field is re-synthesized at the ear canal entrances of the listener. This is in contrast to, for example, stereophony, where a sound field is created that is perceived similar to a natural sound field by humans but that has a physical structure that can depart significantly from that of a natural field.

A physically accurate representation of a sound field cannot be obtained from a real world microphone array over the entire audible frequency range (Meyer and Elko, 2002; Rafaely, 2005). It has been difficult to anticipate perception of the rendered signals based on an instrumental analysis of the signal properties. A number of studies are available in the literature that aim at filling this knowledge gap.

The studies presented in Avni *et al.* (2013), Melchior *et al.* (2009), Neidhardt (2015), Nowak *et al.* (2016), and Nowak and Klockgether (2017) investigate—and some of them also predict—the perception with respect to overall quality or with respect to higher-level attributes that were either elicited from the subjects themselves or prescribed by the experimenter. Array captures with different parameters were compared to each other.

The studies performed in Ahrens *et al.* (2017), Andersson (2017), Bernschütz (2016), and Neidhardt (2015) compared headtracked headphone renderings of array recordings to headtracked headphone renderings of dummy head (DH) recordings of the same scenarios and thereby allowed for drawing conclusions on the authenticity of the array renderings if the DH auralization is assumed to be the ground truth. The number of studies presented in Bernschütz (2016) is extensive. The work presented in the present article may be considered a complement to these.

The rendering stage excluding the limitations of the capture side were investigated in McKenzie *et al.* (2018) and Zaunschirm *et al.* (2018), and enhancements were proposed and validated. Note that it is not a fundamental requirement that the microphone array is spherical. A method using an arbitrary microphone arrangement together with numerically optimal rendering was presented in Rasumow *et al.* (2013).

The trend that a higher-order rendering leads to a better perceptual result is apparent in most of the mentioned studies although some of the experiment paradigms do not allow for

[a]Electronic mail: jens.ahrens@chalmers.se

drawing conclusions in this regard. The order above with the result does not improve further seems to be around the order of 8 (Ahrens *et al.*, 2017; Bernschütz, 2016) whereas such a threshold is not apparent in all results cited above.

The experiment that we present in this article uses spherical arrays and is an extension of the experiments from Ahrens *et al.* (2017) and Andersson (2017) and evaluates the complete end-to-end signal processing pipeline without enhancements. It aims at an objective evaluation of the renderings, i.e., an evaluation that does not involve an individual internal reference or preference (Letowski, 1989). The initial experiment was presented in Andersson (2017) and comprised a scaling of a total of eight attributes that were inspired by the spatial audio quality inventory (SAQI) (Lindau *et al.*, 2014) and were related to both timbre and spaciousness. Note that SAQI is mostly a spatial character inventory as only few of the attributes relate to a personal preference of the rating individual.

The results of said experiment did not exhibit interpretable tendencies although informal listening suggested that such tendencies are likely to be apparent. Our conclusion was that the subjects were not trained sufficiently, and that the experiment paradigm was too challenging. We therefore simplified the subjects' task considerably and asked them to rate exclusively the perceptual distance of stimuli with respect to spaciousness as the observed differences with respect to timbre were only minor (Ahrens *et al.*, 2017), which proved to be more successful.

In the present experiment, we investigated the perceptual distance with respect to spaciousness and timbre in order to have a more comprehensive coverage of the expected relevant attributes. In the context of the present experiment, we define spaciousness in a broader sense than it is traditionally done in concert hall acoustics (Griesinger, 1996). We subsume all attributes of the stimuli that are related to space such as sound source distance, spatial extent of the sound source, perceived size of the acoustic space, and duration and strength of the reverberation among others under this term. This means that the two attributes that are being investigated—spaciousness and timbre—are compound and multidimensional attributes.

We occasionally use the term *virtual sound source* in this article, which we define as a sound field that is identical to the sound field of an actual sound source in a given domain without the sound source being apparent. A sound source recorded with a microphone array and then rendered over headphones is an example for such a virtual sound source.

The paper is organized as follows. Section II outlines the theory of sound field analysis using spherical microphone arrays and sound field (re-)synthesis over headphones. Section III presents the listening experiment that was conducted, Secs. IV and V discuss the results, and Sec. VI presents concluding remarks.

II. THEORY

This section outlines the theory underlying sound field analysis, in this case the capture and decomposition of sound fields by means of spherical microphone arrays, and subsequent re-synthesis of these sound fields. We emphasize that the relevant literature is vast and that we can therefore not present a complete treatment. We rather present a conceptual overview. A matrix-based notation of the following is available in Zaunschirm *et al.* (2018).

Any interior sound pressure field $S(\vec{x},\omega)$ at a location $\vec{x}$ and at angular frequency $\omega = 2\pi f/\mathrm{c}$, where c denotes the speed of sound, can be described in a domain that is free of sound sources or boundaries by (Williams, 1999)

$$S(\vec{x},\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \breve{S}_n^m(\omega) j_n\left(\frac{\omega}{\mathrm{c}}r\right) Y_n^m(\beta,\alpha). \tag{1}$$

$\breve{S}_n^m(\omega)$ are the spherical harmonics expansion coefficients, $j_n(\cdot)$ is the spherical Bessel function of first kind of order n, r is the radial coordinate in the spherical coordinate system, and $Y_n^m(\beta,\alpha)$ are the spherical harmonics basis functions, which are dependent on the colatitude β and the azimuth α of the point of interest $\vec{x}$. Spherical harmonics are an orthonormal basis for square-integrable functions on the surface of a sphere. We skip an explicit definition here as several slightly different definitions exist, which are conceptually identical but make the following mathematical outline complicated as different cases need to be differentiated for the different definitions. We refer the reader to the various references of this article, in particular to Williams (1999).

The expansion coefficients $\breve{S}_n^m(\omega)$ contain all information on the sound pressure field and can be obtained from a spherical Fourier transform along the surface of a notional sphere with radius R centered around the origin of the coordinate system—which then also constitutes the center of expansion—as

$$\breve{S}_n^m(\omega) = \frac{1}{4\pi \mathrm{i}^n R^2 j_n\left(\frac{\omega}{\mathrm{c}}R\right)} \int_{\Omega} S\left(\vec{x}|_{r=R},\omega\right) Y_n^m(\beta,\alpha)^* dA_{\Omega}, \tag{2}$$

where i denotes the imaginary unit, Ω the surface of the sphere, dA_{Ω} is an infinitesimal surface element on Ω, and the asterisk denotes complex conjugation. The factor R^2 arises because the integration in Eq. (2) is not along a unit sphere.

A. Sound field analysis

Measuring the expansion coefficients $\breve{S}_n^m(\omega)$ via Eq. (2) would require a continuous layer of acoustically transparent pressure microphones arranged along a spherical surface. This implementation exhibits two major drawbacks: (1) it is not feasible in practice and (2) this approach requires multiplying by the term $1/(4\pi \mathrm{i}^n R^2 j_n((\omega/\mathrm{c})R))$, which is termed the *radial filter* in microphone array literature. $j_n((\omega/\mathrm{c})R)$ exhibits zeros so that the coefficients $\breve{S}_n^m(\omega)$ cannot be obtained for certain frequencies.

It has proven favorable to arrange the pressure microphones along the surface of a rigid spherical scattering object (Meyer and Elko, 2002; Rafaely, 2005). The presence of the scattering object obviously alters the microphone signals compared to the free-field case discussed previously.

Fortunately, the scattered sound field can be removed from the data in the spherical harmonics domain by modifying the radial filter as

$$\breve{S}_n^m(\omega) = \overbrace{\frac{1}{4\pi \mathrm{i}^n R^2 \left(j_n\left(\frac{\omega}{\mathrm{c}}R\right) - \frac{j_n'\left(\frac{\omega}{\mathrm{c}}R\right)}{h_n'\left(\frac{\omega}{\mathrm{c}}R\right)} h_n\left(\frac{\omega}{\mathrm{c}}R\right) \right)}}^{=d_n(\omega,R)} \times \int_\Omega S_{\mathrm{tot}}\left(\vec{x}|_{r=R}, \omega\right) Y_n^m(\beta,\alpha)^* dA_\Omega, \tag{3}$$

where R denotes the radius of the scattering object, $j_n'(\cdot)$ the derivative of the spherical Bessel function with respect to the argument, $h_n(\cdot)$ and $h_n'(\cdot)$ are the spherical Hankel function and its derivative, respectively, and $S_{\mathrm{tot}}(\vec{x}|_{r=R}, \omega)$ is the total sound pressure field on the surface of the scattering object and is composed of the incident and the scattered sound field. The radial filter from Eq. (3) does not exhibit poles so that the coefficients $\breve{S}_n^m(\omega)$ can be obtained for all frequencies.

Implementing a continuous layer of microphones is not possible but a finite set of discrete microphones has to be used. The integral in Eq. (3) is therefore approximated by a summation as

$$\breve{S}_n^m(\omega) = d_n(\omega, R) \sum_{\Omega_i} b_i S(\vec{x}_i|_{r=R}, \omega)\, Y_n^m(\beta_i, \alpha_i)^*. \tag{4}$$

The index i runs over the entire set of sampling points. The weights b_i are generally required to maintain orthogonality. Many different sampling grids have been discussed in the literature. We refer the reader to, for example, Rafaely (2005) and Zotter (2009).

The discretization in Eq. (4) has two major consequences: (1) the coefficients can be obtained only up to a certain maximum order $n = N$. This means that the infinite summation in (1) has to be approximated by a finite one. A higher order is equivalent to higher physical accuracy. And (2) spatial aliasing, i.e., spatial ambiguities, arise. Theoretically, spatial aliasing is apparent at any time-frequency. There is a frequency f_A

$$f_A = \frac{N\mathrm{c}}{2\pi R}, \tag{5}$$

above which the aliasing has significant magnitude (Rafaely, 2005). f_A is termed the *spatial aliasing frequency*. Spatial aliasing constitutes ambiguities in the spatial information, but it also affects the time-frequency transfer function.

The gain that is applied by the radial filters in either case (2) or (3) can be very high at low frequencies and also at high frequencies. This is a limitation in practice as the uncorrelated self-noise of the microphones will produce a noisy result when large gains are applied. The gain therefore has to be limited in practice. This is equivalent to an order reduction that those frequencies at which the limit is effective.

B. Sound field synthesis

Once the coefficients $\breve{S}_n^m(\omega)$ of a sound pressure field $S(\vec{x}, \omega)$ are available, $S(\vec{x}, \omega)$ can be synthesized by means of loudspeaker arrays or by means of headphones. Headphone-based synthesis is conceptually identical to loudspeaker-array-based synthesis, whereby the loudspeaker array is virtualized by means of head-related transfer functions in the case of headphone presentation. This was used in, for example, Bernschütz (2016) and Duraiswami *et al.* (2005).

We chose the approach from Avni *et al.* (2013), which does not assume a discrete virtual loudspeaker array but performs the rendering, i.e., the application of the head-related transfer functions (HRTFs), directly in the spherical harmonics domain. Conceptually, this constitutes rendering with a continuous layer of an infinite number of infinitesimal loudspeakers. The advantage compared to the use of a discrete set of virtual loudspeakers is the fact that no spatial aliasing is produced on the rendering side. A significant decrease in perceptual quality can be observed in some variants of the discrete approach (Bernschütz, 2016, Sec. 5.6.1). The aliasing that occurs on the capture side cannot be avoided.

Mathematically, we proceed as follows: Any sound pressure field $S(\vec{x}, \omega)$ can be represented by a continuum of plane waves propagating in all possible directions [Williams, 1999; Ahrens, 2012, Eq. (2.45)]. The strength of each plane have is denoted by the complex coefficient $\bar{S}(\phi, \theta, \omega)$, where ϕ and θ denote the propagation direction of the plane wave under consideration. $S(\vec{x}, \omega)$ is synthesized by integrating all plane waves over all possible propagation directions.

Assuming that HRTFs constitute the acoustic response of the human body to a plane wave then the response to $S(\vec{x}, \omega)$ can be determined by integrating over all possible angles as

$$E_{l,r}(\omega) = \int_\Omega H_{l,r}(\phi, \theta, \omega)\, \bar{S}(\phi, \theta, \omega)\, dA_\Omega, \tag{6}$$

where $E_{l,r}(\omega)$ denotes the ear signal at the left and right ear, respectively, and $H_{l,r}(\phi, \theta, \omega)$ are the left and right HRTFs for propagation direction (ϕ, θ) of the plane wave.

Discretizing the integral in Eq. (6) is equivalent to rendering via a virtual discrete loudspeaker array. This was employed in most of the studies cited in Sec. I.

Expanding all quantities inside the integral into spherical harmonics and exploiting the orthogonality of the spherical harmonics allows for resolving the integral so that the left and right ear signals $E_{l,r}(\omega)$ of a listener with HRTFs $H_{l,r}(\phi, \theta, \omega)$ exposed to the sound field $S(\vec{x}, \omega)$ are given by

$$E_{l,r}(\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(\omega, R) a_m \breve{S}_{n,tot}^{-m}(\omega) \breve{H}_n^m(\omega), \tag{7}$$

where $d_n(\omega, R)$ is the radial filter from Eq. (3), and $\breve{S}_{n,tot}^{-m}(\omega)$ are the coefficients of $S_{\mathrm{tot}}(\vec{x}|_{r=R}, \omega)$, the sound pressure on the surface of the rigid sphere. $\breve{H}_n^m(\omega)$ are the expansion coefficients of $H_{l,r}(\phi, \theta, \omega)$. The factor a_m in Eq. (7) depends on the definition of the spherical harmonics that is used. We refer the reader to (Andersson, 2017, Sec. 2.4) for details.

In summary, the low-frequency performance of spherical microphone arrays is limited by sensor self-noise and sensor placement errors, which can cause large errors due to the partially large gains of the radial filters. The high-frequency performance is limited by the finiteness of the amount of sensors used. Feasible practical arrays are physically accurate over a bandwidth of around 2 octaves, say, between approximately 500 and 2000 Hz.

Note that the head orientation of the listener is encoded in the HRTFs $H_{l,r}(\phi, \theta, \omega)$. Either the sound field or the HRTFs can be rotated to compute the ear signals for different head orientations. Particularly head orientations about the vertical axis by an arbitrary angle α_{rot} are straightforward to implement by adding a factor $e^{-im\alpha_{\text{rot}}}$ in Eq. (7). These are the head orientations that we track in the listening experiment presented below.

III. LISTENING EXPERIMENT

A. Stimulus preparation

All stimuli were produced based on the data from Stade *et al.* (2012). These data comprise impulse response measurements in different rooms from a loudspeaker to the microphones of different rigid-sphere arrays as well as to the ears of a DH in the same location as the microphone array. The measurements were performed such that time invariance of the rooms may be assumed. The signal processing was performed in Python using the port and extension of Hohnerlein and Ahrens (2017) of the SOFiA sound field analysis toolbox (Bernschütz *et al.*, 2011).

The DH model was a torsoless Neumann KU100, and the binaural room impulse responses were measured for different head orientations for a complete circle with increments of 1°. Additionally, anechoic head-related impulse response of the DH are provided for a 2702-node Lebedev grid (Bernschütz, 2013). We tested different strategies for computing the spherical harmonics coefficients $\breve{H}_n^m(\omega)$ of the HRTFs $H_{l,r}(\phi, \theta, \omega)$ used in Eq. (7) including expansion of the time-domain data and separate expansions for the magnitude and unwrapped phase of the data in frequency domain. We finally chose expansion of the complex data in frequency domain as the approach proved to be most robust in terms of the sanity of the results. Refer to Andersson (2017) for details.

The different spherical microphone arrays were emulated through the *VariSphear* single-microphone scanning array (Bernschütz *et al.*, 2009). It has a robotic arm that is equipped with a measurement microphone that is flush mounted in a rigid spherical scattering object of a radius of $R = 8.75$ cm. The construction rotates such that the microphone can be moved to arbitrary positions on the surface of the spherical scattering object while keeping the center of the scattering object still. This way, arbitrary sampling grids can be emulated. The data from Stade *et al.* (2012) use a Lebedev grid with different numbers of sampling points. Table I lists the 2 sampling grids that we used.

We used the data for the room Control Room 1 (CR1), an acoustically dry control room of a recording studio with a reverb decay time (−60 dB) of approximately 0.2 s, as well as the room Small Broadcasting Studio (SBS), which is a chamber music recording facility with a reverb decay time of approximately 1 s.

TABLE I. Number of microphone positions used for the different orders.

Order used	No. of microphones	f_A
1	50	3.1 kHz
3	50	3.1 kHz
3[a]	110[a]	5.0 kHz
5	50	3.1 kHz
8	110	5.0 kHz

[a]This configuration is only used for the purpose of comparing two different grids at the same order. Unless specified as different, the third-order stimuli use the 50-microphone array.

The stimuli preparation was performed exclusively based on impulse responses. The input data to the array signal processing pipeline were the room impulse responses of the individual microphones of the arrays as well as the (anechoic) head-related impulse responses of the DH. The output of the processing pipeline was a pair of ear impulse responses that represent the transfer function of the complete pipeline for a given head orientation, i.e., the path of the signal from the loudspeaker through the array and through the rendering stage that virtually puts the DH into the sound field. For each condition, 360 pairs of ear impulse responses were computed representing 360 head orientations whereby the rotation occurs about the vertical axis through the head center. Similarly, the ground truth, i.e., the direct DH measurements of the rooms, were also available for 360 different head orientations.

The fact that all processing was performed based on impulse responses means that the microphone signals are free of additive noise such as sensor self-noise. The fact that the same single microphone was used for all array measurements means that the data are free of microphone mismatch. The presented listening experiment has therefore been performed under ideal conditions. We used a gain limit for the radial filters of 0 dB, which is on the conservative side. We chose this gain limit to make the listening experiment compatible with ones from Bernschütz (2016).

The presentation of the stimuli was performed using the software SoundScape Renderer (SSR) (Geier and Spors, 2010; Geier *et al.*, 2008) running in binaural room synthesis mode. SSR convolves a given input signal with the pair of impulse responses that corresponds to the instantaneous head orientation as provided by a headtracker. The use of headtracking is essential in such studies in order to avoid distortion of the spatial perception (Begault *et al.*, 2000; Lindau, 2014). We employed a Polhemus Patriot headtracker.

We chose an acoustically dry rock drum recording with a duration of 90 s as stimulus. The quarter notes of the drum rhythm occurred at approximately 180 bpm, which is a tempo that is excitatory but does not feel rushing. Drums are a very critical signal in that they contain strong transients as well as they exhibit a broad spectrum. Preliminary experiments with less critical signals such as speech produced only minuscule differences between different conditions so that we

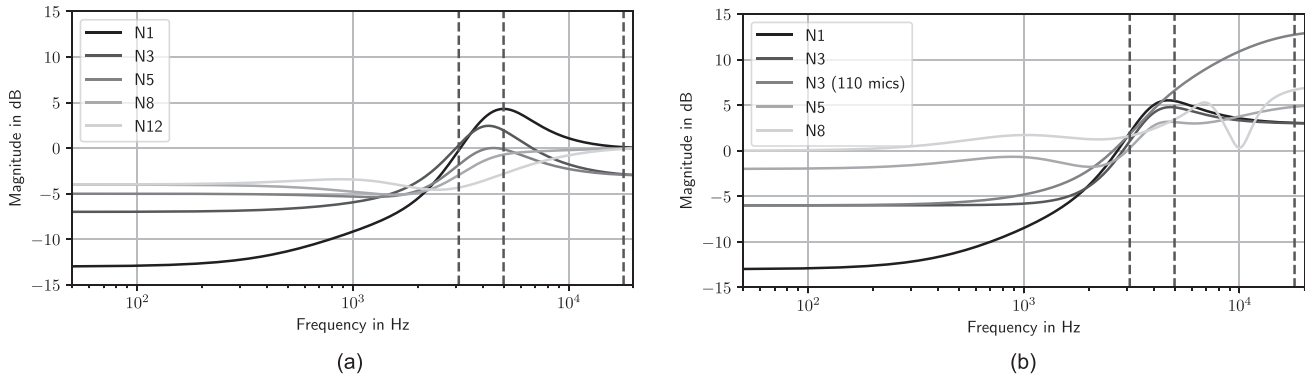

FIG. 1. Equalization curves for the different orders for (a) CR1 and (b) SBS; f_A from Eq. (5) is also indicated.

chose to investigate only the critical case. Similar observations are reported in Bernschütz (2016). The effect of the source signal on perception for other less critical signals was found to be low or not existent (Bernschütz, 2016).

When a change in the head orientation occurs, then SSR convolves the current signal block with the current as well as with the previous set of filters and crossfades with a cosine ramp between the signals. The block size was set to 256 samples at a sampling frequency of 48 kHz with 2 blocks of buffering. The overall latency of the system is composed of the latency of the tracker (18.5 ms), 2 blocks of buffering (2×5 ms), 1 block delay due to signal routing and processing (5 ms), and approximately a half block delay due to the crossfade after the convolution (2.5 ms). This amounts to 36 ms, which is well below audibility (Lindau, 2009).

We used a pair of AKG K702 open-design headphones for the experiment. We used the minimum-phase compensation filter for this headphone model that is provided with the data set (Stade *et al.*, 2012) to compensate for the headphones' transfer function.

It was shown in Avni *et al.* (2013) and at other locations in the literature that the order limitation of a sound field has a noticeable effect on the magnitude transfer function of the overall system, which can lead to coloration. Automatic equalization strategies have been presented for the frequency range below the aliasing frequency (Ben-Hur *et al.*, 2017). Automatic equalization above the aliasing frequency for the present scenario is unsolved [equalization of the rendering stage is presented in McKenzie *et al.* (2018)]. We chose to perform manual equalization in the following manner.

We used the direct DH measurement data for the sound source being straight ahead as reference. Then, the corresponding transfer function of the array pipeline was equalized by hand using a series of a 2nd order low-shelving filter, a varying number of 2nd order peak/notch filters, and a 2nd order high-shelving filter such that the difference in timbre becomes minimal. The same filter set with the same parameters were then applied to all other head orientations in the array pipeline. This may be termed a global equalization. Two peak-filters were used for most of the stimuli. Only the stimulus of order 5 for SBS used three peak/notch-filters, and the stimulus of order 8 for SBS used four peak/notch-filters.

The resulting equalization curves are illustrated in Figs. 1(a) and 1(b), and a comparison of the transfer function of the DH for the orientation straight ahead in room CR1 with a corresponding equalized example output of the microphone array pipeline is depicted in Fig. 2.

We incorporated also lowpassed versions of some of the stimuli in the experiment. The motivation was twofold: (1)

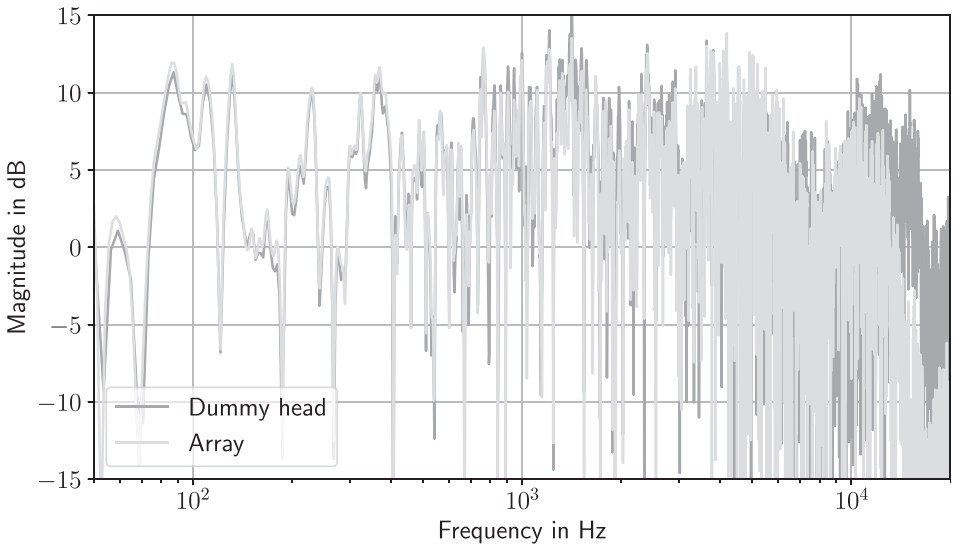

FIG. 2. Example result of the manual equalization for the binaural transfer function of room CR1 to the left ear; the array employed 50 microphones and order 5.

lowpassing attenuates the frequency range in which spatial aliasing occurs and thereby allows to suppress its effect and (2) lowpassing produces a timbral difference that is easy for the subjects to detect. This gives the subjects positive feedback regarding their ability to perform the required task as many stimulus conditions produce only very small timbral differences. We chose a maximum flat second order lowpass filter with a cutoff frequency of 3000 Hz and quality factor of $Q = 1/\sqrt{2}$.

We also found in preliminary studies (Ahrens *et al.*, 2017; Andersson, 2017) that the perceptual order-dependency of the auralization is different for virtual sound sources in front of the listener compared to lateral virtual sound sources. We therefore included also conditions in which we rotated all data sets by 90° so that the listener hears the sound source fully lateralized to the left. We chose to rotate only in one direction to avoid uncertainty that can arise due to possible asymmetries in the auditory system (Blauert, 1997). This caused a slight direction imbalance for the subjects as sound sources appeared only in front or to the left. The subjects did not report this to be irritating.

B. Experiment paradigm

Assuming ideal conditions, then the signals that are produced by the spherical microphone array pipeline are identical to the signals that arise at the ears of a listener who is exposed to the sound field that was captured by the array. This requires a continuous distribution of perfectly matched and noiseless pressure microphones along the surface of the spherical scattering object as well as that the HRTFs of the listener are being employed in the rendering stage. Any departure of the output of the processing pipeline from the actual ear signals may be interpreted as artifacts.

In the present study, the HRTFs that we employed in the rendering are the HRTFs of the same DH with which the room responses were measured. We may therefore interpret the direct DH measurement data as ground truth against the output of the processing pipeline is being compared. We chose an A-B-comparison with attribute scaling as experiment paradigm as the observed differences between the reference and a stimulus can be very small (Bech and Zacharov, 2006). Most of the times, the subjects perceptually compare the output of the processing pipeline for different parameter sets against a direct auralization of the DH data in the same room (and at the same location). We added a few conditions in which the comparison is not performed against the DH data:

- A hidden reference (DH vs DH)—to assess the reliability of the subjects' responses.
- Renderings of the same order but obtained from arrays with different numbers of microphones—to assess the influence of the sampling grid.
- Non-lowpassed stimuli vs lowpassed—see below for the motivation.

Preliminary experiments showed that the timbre produced by the array pipeline can be very similar to the timbre of the DH data. This makes the task of rating the magnitude of the difference difficult for the subjects. If an experiment contains only pairs of stimuli that produce only minuscule differences, then this can demotivate the subjects as they can lose confidence in their ability to perform the required task. We therefore added a few pairs of non-lowpassed stimuli vs lowpassed stimuli—which produce a significant difference regarding timbre—to assure that there are cases in which the subjects find confirmation that they master the task.

The sound source that is being rendered is located in the horizontal plane in all cases. We do therefore not assume that the employment of non-individual HRTFs limits the validity of the results.

Previous studies including Ahrens *et al.* (2017) suggest that acoustically dry rooms require higher orders to be perceptually satisfying when rendered over headphones. To account for this while keeping the amount of conditions minimal, we employed different sets of orders for the different rooms. Table II provides a complete list of stimulus pairs.

C. Procedure

The subjects were seated in front of a computer screen with a keyboard and a mouse in a quiet room. Their task was to rate the perceived difference between the stimuli (1) with respect to spaciousness by moving a slider along a continuous scale ranging from "stimulus A is a lot more spacious than stimulus B" to "stimulus B is a lot more spacious than stimulus A", as well as (2) with respect to timbre, whereby in this case, only the magnitude of the difference was to be rated by moving a slider along a continuous scale ranging from "no difference" to "huge difference." The graphical user interface (GUI) is depicted in Fig. 3.

TABLE II. List of stimuli pairs that were tested; the button assignment (A and B) in the graphical user interface was randomized; NX refers to a rendering of Xth order; (lp) refers to a lowpassed stimulus; all stimuli pairs were presented non-rotated as well as rotated unless specified as different.

Room	Stimulus 1	Stimulus 2
CR1	DH	N1
CR1	DH	N3
CR1	DH	N5
CR1	DH	N8
CR1	DH	N12
CR1	DH	DH (lp)
CR1	DH	N1 (lp)
CR1	DH (lp)	N5 (lp)
CR1	N8	N8 (lp)
SBS	DH	N1
SBS	DH	N3
SBS	DH	N5
SBS	DH	N8
SBS	N3 (50 nodes)	N3 (110 nodes)
SBS	DH	DH (lp)
SBS	DH	N1 (lp)
SBS	DH (lp)	N5 (lp)
SBS	N8	N8 (lp)
SBS	DH[a]	DH[b]

[a]This hidden-reference condition was presented only for the listener virtually facing the sound source (i.e., not rotated).

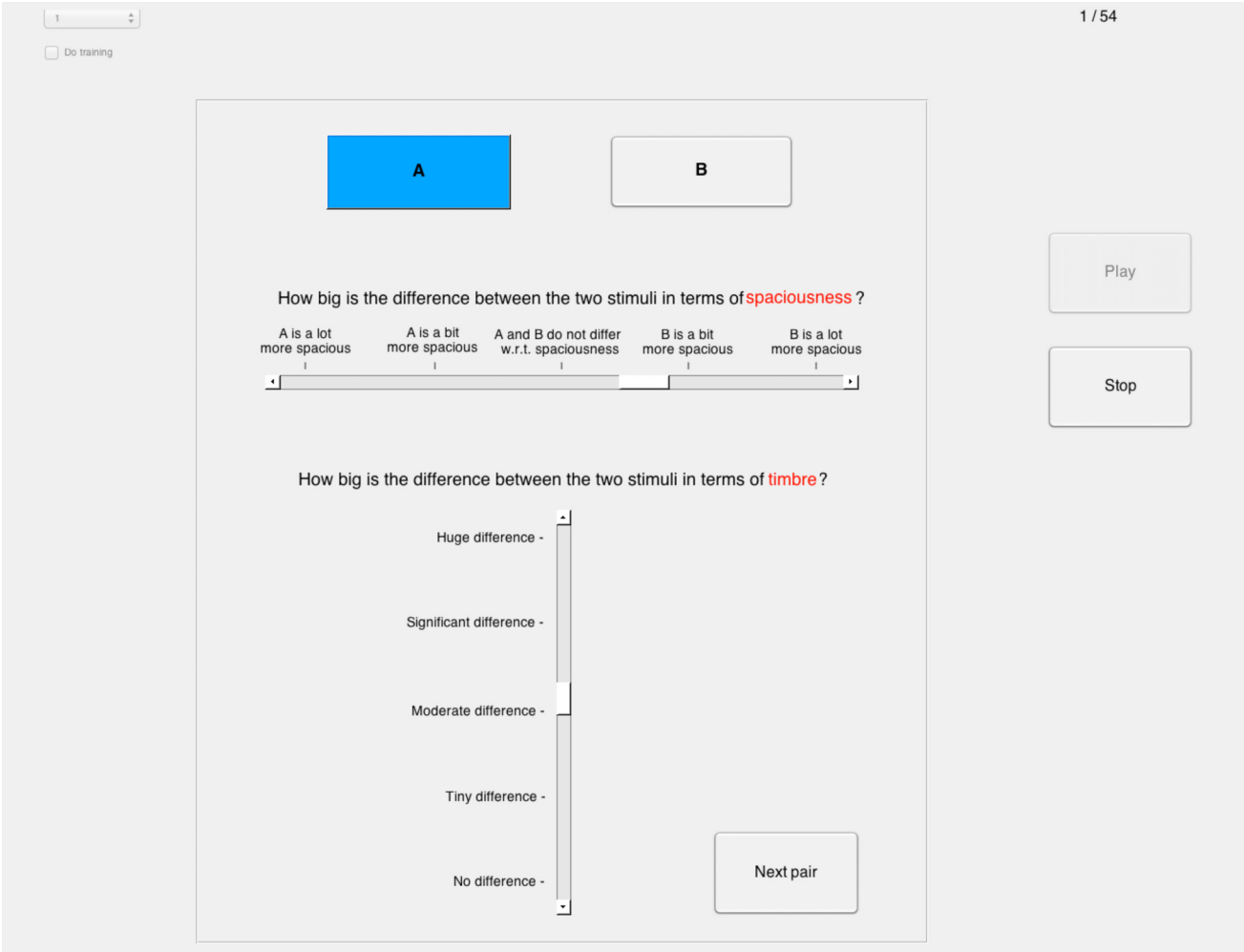

FIG. 3. (Color online) The GUI of the experiment.

The subjects were instructed to rate a stimulus to be more spacious than another one if one or more of the following differences occur. (1) The sound source sounds farther away. (2) The room sounds larger. (3) The reverberation sounds stronger.

They were also instructed to give a random rating if contradictions with respect to above differences occur.

Switching between the stimuli of a given stimulus pair was possible either through mouse clicks on the corresponding buttons or via keys on the keyboard. The handles of the slider always appeared in the neutral position for each new stimulus pair (i.e., "A and B do not differ with respect to spaciousness" and "no difference," respectively).

The subjects were made aware of the fact that head-tracking was employed but they were not specifically instructed to make conscious use of this feature.

20 subjects of both male (70%) and female (30%) gender participated in the experiment. The age range was between 24 and 40 years with a median of 29 years. The subjects participated voluntarily and were recruited from students of the Sound and Vibration Master's program at Chalmers University of Technology as well as from staff members of the Division of Applied Acoustics. Most subjects did not have prior experience with listening experiments. The experiment was divided into two sessions, which occurred on different days. Only one room was tested in any given session, whereby the order of rooms was randomized. Each session started with written instructions and a set of 6 pairs of stimuli for training.

Each stimulus pair was presented three times in total in randomized order and with randomized button assignment ("A" and "B"). This results in 54 stimuli for room CR1 and 57 stimuli for room SBS.

During the experiment, the drum loop was playing continuously without interruptions or pauses. The subjects were able to monitor their progress during the session by means of a stimulus-pair counter (cf. top-right in Fig. 3).

IV. RESULTS

All 20 subjects produced consistent responses so that all recorded data are considered in the following. We recorded a total of $(54+57)\times 20=2220$ ratings for each spaciousness and timbre. All subjects were interviewed after each session and were asked to provide comments about the experience. All subjects confirmed that they were feeling confident with the task. No prominent artifacts in the signals were reported. Note that the subjects in Avni *et al.* (2013) identified unpleasant artifacts in the synthetic data that were used there.

The duration of the actual experiment segment ranged from 10 to 45 min for each session with a mean of 25 min.

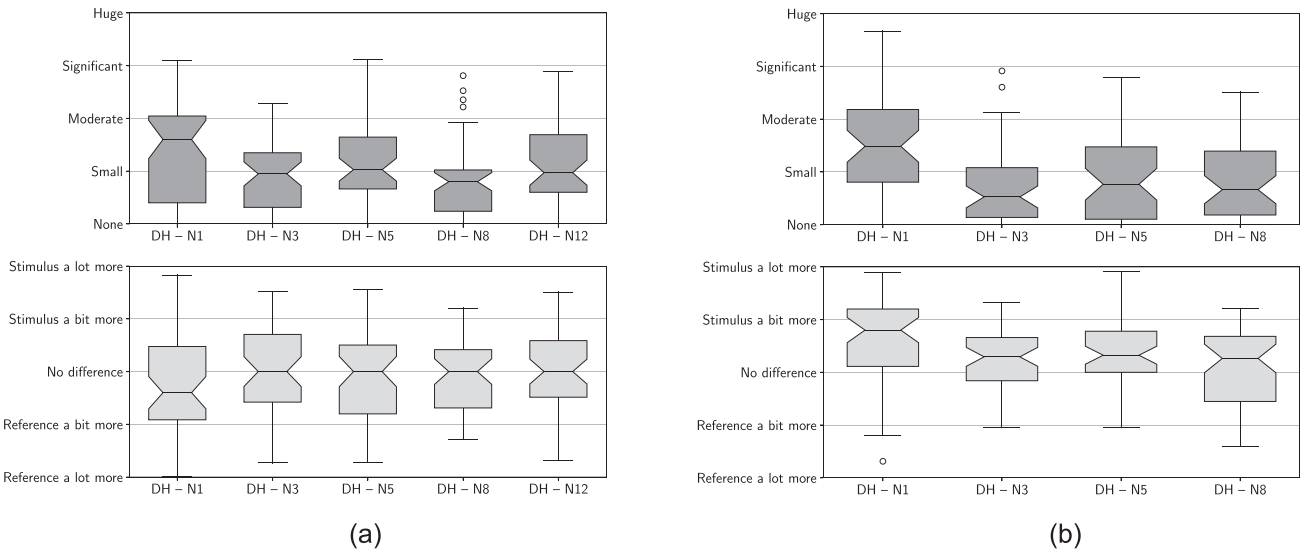

FIG. 4. Perceived difference for a frontal sound source; top: timbre; bottom: spaciousness; (a) CR1; (b) SBS.

This corresponds to 12, 51, and 29 s per stimulus pair, respectively. No dependency on the room was observed.

The subjects performed only small conscious head movements.

The results are presented as boxplots in Figs. 4–8. The boxplots show the median value of the data via the horizontal line, the 25th and 75th percentiles via the gray box, the whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually via circles. The notches represent the 95% confidence interval of the median. The top row with dark gray boxes represents the ratings of the difference with respect to timbre, and the bottom row with light gray boxes represents the difference ratings with respect to spaciousness.

A. Statistical analysis

A statistical analysis was performed on the data for both rooms for orders 1, 3, 5, and 8 for the non-lowpassed stimuli. The Anderson-Darling test showed that the data can be assumed normally distributed only for some of the test conditions (Anderson, 1954). In order to identify inconsistent ratings, we computed the variance of all responses of one subject for each condition. Values larger than $Q3 + 3 \times IQR$, with Q3 being the 3rd Quartile and IQR being the Inter Quartile Range, were considered outliers and were disregarded in the subsequent analysis. This resulted in 26 data points being disregarded (15 for timbre and 11 for spaciousness).

We also disregarded the polarity of the rating scale for spaciousness in the subsequent analysis, i.e., we assumed the scale to range from "no difference with respect to spaciousness" to "a lot of difference with respect to spaciousness." This made the rating scale compatible with the scale for timbre ("no difference" to "huge difference"). Both scales were then scaled to range from 0 to 1 in numerical terms.

We applied a linear mixed model with the subject as random factor (Bortz, 2006). Table III lists the results for the main effects and all statistically significant interaction effects.

It is evident from Table III that the array order and the listener orientation are significant ($p < 0.05$), whereas the room and the subject are not significant variables ($p \geq 0.05$).

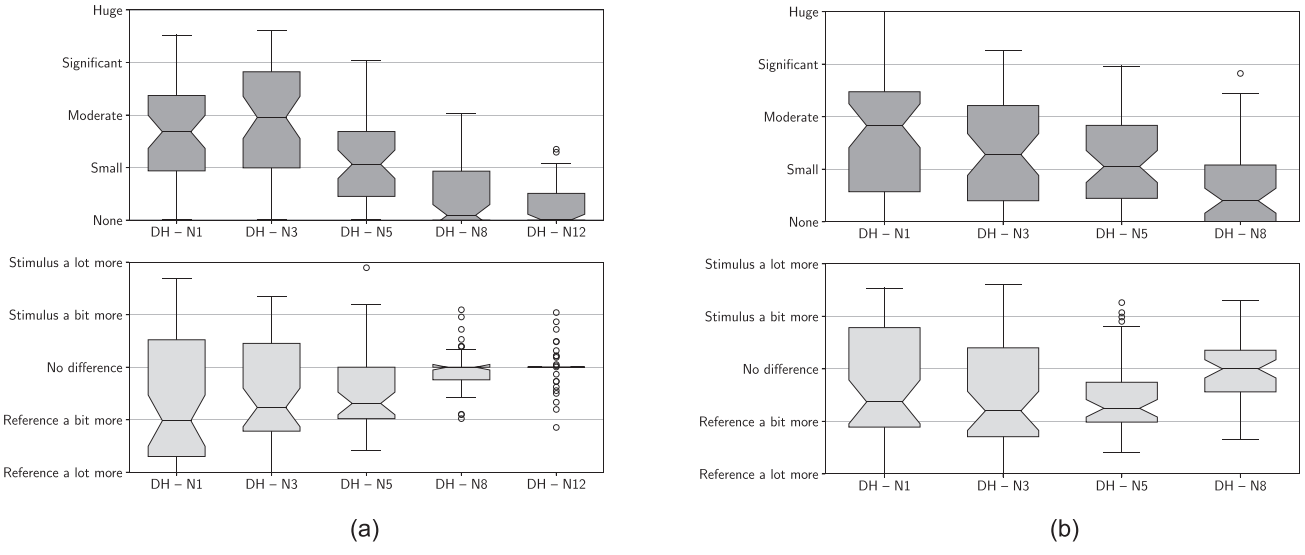

FIG. 5. Perceived difference for a lateral sound source; top: timbre; bottom: spaciousness (a) CR1; (b) SBS.

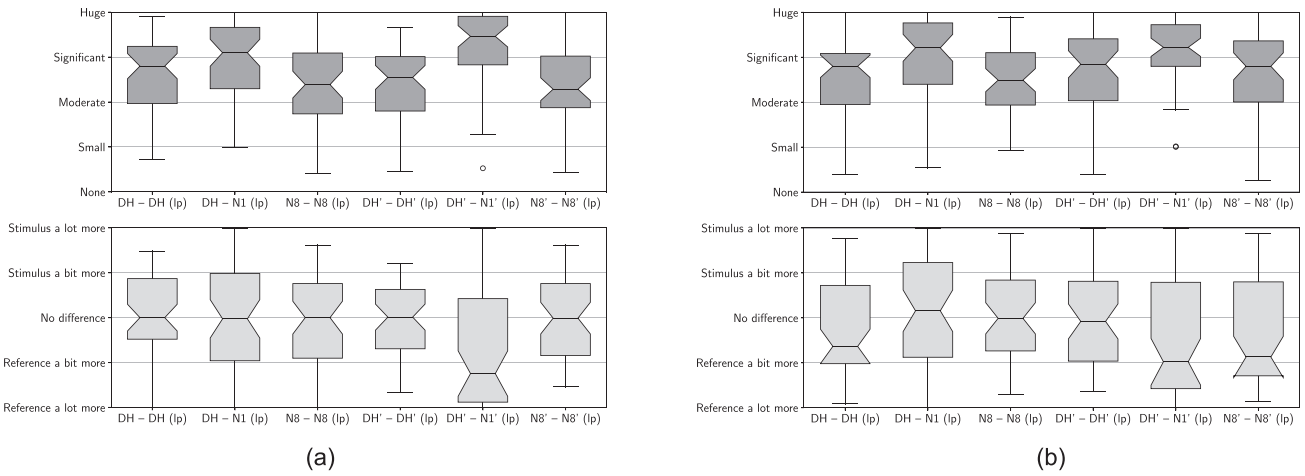

FIG. 6. Perceived difference; the reference is not lowpassed (lp); the primed stimuli comprise the lateral sound source; top: timbre; bottom: spaciousness; (a) lowpassed CR1; (b) lowpassed SBS.

We also listed the significant interaction effects. The interaction room–array order indicates that there is a room-dependent difference in how the responses change with the order. The interaction array order–listener orientation represents the circumstance that the effect of the array order depends strongly on the listener orientation, which is evident when comparing Figs. 4 and 5 and which is discussed in detail below.

B. Further observations for room CR1

1. Frontal virtual sound source position

The order dependency of the ratings for those pairs of stimuli in which the virtual source was located in front of the listener is moderate as depicted in Fig. 4(a). No difference in terms of spaciousness is perceived on average, whereby the variance is high. The perceived difference in terms of timbre ranges on average around "small" with high variance.

2. Lateral virtual sound source position

The situation is different for the case when the virtual sound source is located to the left side of the listener as depicted in Fig. 5(a). For low orders, the reference, i.e., the direct DH auralization, is rated more spacious. This bias decreases towards higher orders and vanishes completely for the orders 8 and 12. Similarly, the perceived difference in terms of timbre is in the range between small and significant for low orders and vanishes completely for 12th order.

3. Influence of the lowpass filtering

The reported difference in terms of timbre is consistently high for all cases in which a lowpassed stimulus was compared with an non-lowpassed stimulus [cases DH-DH (lp), DH-N1 (lp), N8-N8 (lp), and the according rotated conditions in Fig. 6(a)]. This is expected. Recall that these stimulus pairs are only added to keep the subjects motivated as

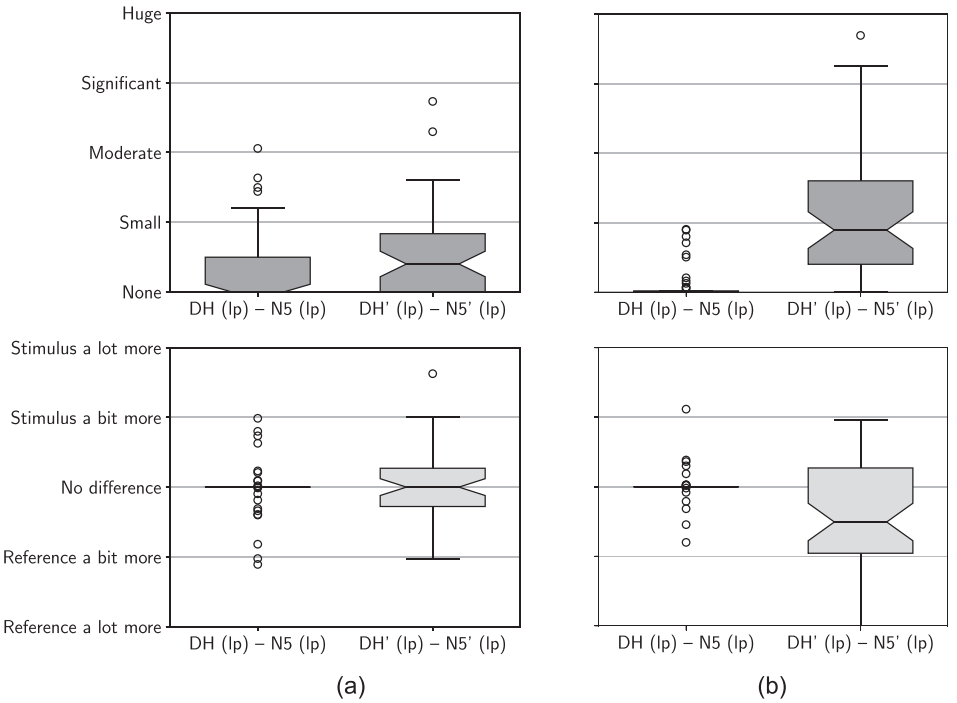

FIG. 7. Perceived difference; all stimuli are lowpassed (lp); the primed stimuli comprise the lateral sound source; top: timbre; bottom: spaciousness; (a) lowpassed CR1; (b) lowpassed SBS.

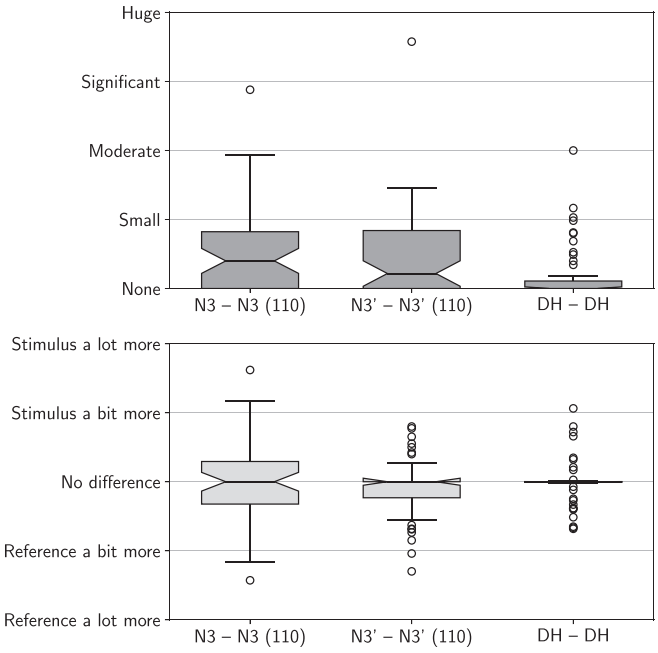

FIG. 8. Room SBS, special cases; the primed stimuli comprise the lateral sound source; top: timbre; bottom: spaciousness.

explained in Sec. III B. No differences with respect to spaciousness are reported on average. The variance is very high in all cases.

The cases where both stimuli were lowpassed are depicted in Fig. 7(a). No difference was reported for the case DH (lp)-N5 (lp) for the frontal source position and only a minor difference in terms of timbre was reported for the lateral source position.

C. Further observations for room SBS

1. General

The statistical analysis in Sec. IV A confirmed that the room has no significant influence on the responses. The observations from Sec. IV B therefore also hold for room as SBS. This is also as evident from Figs. 4(b), 5(b), 6(b), and 7(b): For frontal virtual sound source positions, the reported perceived difference shows only a minor dependency on the order of the rendering. The reported difference decreases significantly towards higher orders for the lateral virtual sound source position and is very low for the highest tested order of 8. No difference is perceived if both stimuli were lowpass filtered and the source was in front; a small difference is perceived when the virtual source is lateral.

2. Special cases

The hidden reference, i.e., the pair DH-DH, was identified reliably as evident from Fig. 8

The perceived difference between a 3rd-order rendering of the scenario when captured with 50 microphones and a 3rd-order rendering of the same scenario when captured with 110 microphones is zero with respect to spaciousness and between "none" and "small" with respect to timbre.

V. DISCUSSION

Many of the ratings exhibit a high or very high variance. Interestingly, we observed a rather low intrasubject variance, i.e., the responses of a given subject to any given stimulus pair are very consistent for all three occurrences of that stimulus pair. However, the intersubject variance is high, i.e., different subjects respond differently to the same stimulus pair. The reason for this could be the fact that the attributes to be scaled are multidimensional an each subject uses a different weighting for the individual dimensions. See also further comments in Sec. V D.

Our results confirm many observations from Bernschütz (2016). The following discussion focuses on the new aspects.

A. Order dependency

The dependency of the perception on the spherical harmonics order of the rendering depends on the location of the virtual sound source. The array renderings were rated less spacious than the direct DH auralizations for low orders for lateral virtual sound sources. This is due to the fact that the array renderings produce a virtual sound source that is poorly externalized and sounds closer to the ear than the virtual sound source in the DH auralizations. As mentioned in Sec. III C, the subjects were instructed to rate stimuli in which the sound source appears closer to be less spacious.

The virtual sound source in the array renderings sounds farther away with increasing order and sounds as far away as the source due to the DH for orders 8 and higher. The situation is different for frontal sound sources where no obvious dependency of the ratings on the order is apparent. Comments by the subjects and informal listening by the authors suggest that there is always some sort of difference between the array output and the DH apparent that is difficult

TABLE III. Results of the statistical analysis [$F(a, b)$ means that the variable exhibits a degrees of freedom and that the error is b].

Dependent variable	Results (timbre)	Results (spaciousness)
Room	$F(1, 16.2) = 0.38$, $p = 0.546$	$F(1, 16.0) = 0.01$, $p = 0.923$
Array order	$F(3, 49.7) = 39.5$, $p < 0.001$[a]	$F(3, 48.6) = 37.2$, $p < 0.001$[a]
Listener orientation	$F(1, 16.2) = 4.87$, $p = 0.042$[a]	$F(1, 16.1) = 7.14$, $p = 0.017$[a]
Subject	$F(19, 17.4) = 2.13$, $p = 0.065$	$F(19, 13.6) = 1.45$, $p = 0.247$
Interaction room—array order	$F(3, 49.3) = 6.27$, $p = 0.001$[a]	$F(3, 48.3) = 2.90$, $p = 0.045$[a]
Interaction room—subject	$F(19, 16.2) = 4.87$, $p = 0.001$[a]	$F(19, 20.8) = 1.48$, $p = 0.200$
Interaction array order—listener orientation	$F(3, 48.3) = 17.4$, $p < 0.001$[a]	$F(3, 47.3) = 15.2$, $p < 0.001$[a]

[a]Results that are significant based on a 5% significance level.

to pinpoint. Increasing the order changes the difference without changing its character. We confirmed this observation via informal listening to stimuli of up to 29th order.

An explanation for this source location dependency could be the circumstance that the human auditory system has higher spatial resolution for sound sources in front of the listener and is thereby able to detect the difference between the DH signals and the array auralizations more reliably (Blauert, 1997). The fact that both ears are illuminated without head shadowing for frontal sound sources might increase the resolution of timbre perception compared to lateral sound sources where the energy arriving at the contralateral ear is attenuated significantly.

Although lower orders of, say, 3 seem to be sufficient for rendering sources in front, it is important to realize that head rotations of the listener can evoke substantial changes both with respect to timbre and with respect to spaciousness at such orders. We have not observed any changes with head rotations at orders 8 and 12 in informal listening.

B. Frequency dependency

The subjects reported minor or no differences for those cases where both the DH signals as well as the array signals where lowpassed, cf. Figs. 6 and 7. This supports the conclusion that spatial aliasing is the cause for the perceived differences, as the lowpassing strongly attenuates the frequency range in which aliasing occurs.

Interestingly, the rather conservative gain limitation that we applied to the radial filters does not seem to be audible. As the gain limitation is essentially an order limitation at low frequencies, it seems that the chosen parameter set provides sufficient spatial information for the human auditory system to be indistinguishable from the ground truth.

C. Room dependency

Similar to Bernschütz (2016), we did not observe a significant difference of the ratings between rooms. This is contrary to our previous results from Ahrens *et al.* (2017), where the room SBS required only fifth order to be almost indistinguishable from the DH auralization whereby room CR1 required eighth order. We have no explanation for the differences in the observations. We only tested spaciousness in Ahrens *et al.* (2017), and subjects had to ignore any other differences, which might have made them more tolerant than in the present study.

D. Other

The results show that a larger difference in timbre comes with a larger variance of the ratings of spaciousness. One explanation could be that differences with respect to timbre happen to occur together with differences with respect to spaciousness. Another explanation could be that a difference in timbre can cause a difference in spaciousness and vice versa as these attributes are not orthogonal. Interestingly, our subjects reported an increase of the source distance due to lowpassing—suggesting higher spaciousness—but a reduction in the *presence* of the source—suggesting lower spaciousness as the source sounds smaller. Such contradictions can be an explanation for the large intersubject variance that we observed as each subject might have had a different strategy for dealing with such contradictions.

Another observation is that we found it easier to equalize higher-order renderings compared to lower-order ones, which might also have contributed to the variance for both timbre and spaciousness ratings. Although we occasionally used more filters to equalize higher-order renderings, we used less aggressive settings in this case.

VI. CONCLUSIONS

We presented a listening experiment in which subjects compared auralizations of spherical microphone data with dummy head recordings of the same scenarios with head-tracking. The presented experiment fills a gap that has been apparent in the existing literature between studies that compare array renderings to dummy-head-based ground truth such as Bernschütz (2016) with respect to overall quality and works that investigated different higher-level attributes without a dummy-head-based ground truth such as Avni *et al.* (2013), Nowak *et al.* (2016), and Nowak and Klockgether (2017). The authenticity of the auralization cannot be investigated without a ground truth.

Our experiment determined the perceptual distance between array-based and dummy-head-based auralization depending on the spherical harmonics order. Although we only investigated timbre and a broad interpretation of spaciousness, we have not found indications that other differences can be observed so that these two multidimensional categories can be assumed to be comprehensive.

Our results show that the perceptual differences mostly decrease in magnitude up to an order of 8 above which no further improvement is expected, which confirms the results from Bernschütz (2016). Order 8 requires 110 microphones when using a Lebedev grid, which is at the limit but still feasible in practice. A noticeable yet small difference remains for frontal sound sources whereas the dummy-head-auralization and the array-auralization are indistinguishable for lateral sound sources at such high orders.

Our experiment confirmed the observation that the location of virtual sound source has a significant effect on the perception at lower orders. We have initially reported this in Ahrens *et al.* (2017) and Andersson (2017). To the best of the authors' knowledge, Neidhardt (2015) is the only other study in which a lateral virtual sound source was tested explicitly. The differences were not as prominent as in the present results. This is likely due to the fact that simulated data was used in Neidhardt (2015), which can exhibit limited plausibility and perceptual fidelity as we observed during the work that was published in Avni *et al.* (2013) in which the main author of the present paper was involved. This assumption is supported by the fact that the ratings for overall quality in Neidhardt (2015) tended to be low.

We confirmed the result from Bernschütz (2016) that the employed sampling grid as well as the type of room that is auralized do not have a significant effect.

The investigation of lowpassed stimuli shows that audible differences between dummy head and array occur only at high frequencies as the differences vanish already at the

tested order of 5. The perceptual differences remaining with non-lowpassed stimuli at higher orders may therefore be attributed to spatial aliasing. A comparable result was obtained in Bernschütz (2016).

Our experimental paradigm did not allow for investigating the effect of the self-noise of the microphones. The fact that the conservative gain limitation that we used with the radial filters does not produce a perceptual impairment is a promising result as this setting avoids making the processing pipeline vulnerable to noise and microphone mismatch. We are currently extending the implementation of the processing pipeline to streamed signals, which will then allow for evaluating the impact of additive noise.

Our processing pipeline may be considered the most basic pipeline that yields near-to-authentic results. We chose this setup to document basic performance of this type of auralization. Several components can be tuned and improved. The evaluation of this is subject to future work. Some promising initial results on some aspects are available, for example, on the reduction of spatial aliasing (Alon *et al.*, 2015; Bernschütz, 2016) and on enhancement of the rendering stage (McKenzie *et al.*, 2018; Zaunschirm *et al.*, 2018).

ACKNOWLEDGMENTS

We thank all our subjects of their voluntary participation and Ina Wechsung for advice on the statistical analysis.

Ahrens, J. (**2012**). *Analytic Methods of Sound Field Synthesis* (Springer, Berlin).
Ahrens, J., Hohnerlein, C., and Andersson, C. (**2017**). "Auralization of acoustic spaces based on spherical microphone array recordings," in *Proceedings of Acoustics'17*, ASA/EAA, Boston, MA, pp. 1–4.
Alon, D. L., Sheaffer, J., and Rafaely, B. (**2015**). "Plane-wave decomposition with aliasing cancellation for binaural sound reproduction," in *139th Convention of the AES*, AES, New York, NY, p. 9449.
Anderson, T. W., and Darling, D. A. (**1954**). "A test of goodness-of-fit," J. Am. Stat. Assoc. **49**, 765–769.
Andersson, C. (**2017**). "Headphone auralization of acoustic spaces recorded with spherical microphone arrays," Master's thesis, Chalmers University of Technology, Göteborg, Sweden.
Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., and Rafaely, B. (**2013**). "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," J. Acoust. Soc. Am. **133**(5), 2711–2721.
Bech, S., and Zacharov, N. (**2006**). *Perceptual Audio Evaluation—Theory, Method and Application* (Wiley, Chichester, UK).
Begault, D. R., Lee, A. S., Wenzel, E. M., and Anderson, M. R. (**2000**). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," in *108th Convention of the AES*, p. 5134.
Ben-Hur, Z., Brinkmann, F., Sheaffer, J., Weinzierl, S., and Rafaely, B. (**2017**). "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," J. Acoust. Soc. Am. **141**(6), 4087–4096.
Bernschütz, B. (**2013**). "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of AIA/DAGA*, DEGA, Meran, Italy, pp. 592–595.
Bernschütz, B. (**2016**). "Microphone arrays and sound field decomposition for dynamic binaural recording," Ph.D. thesis, Technische Universität Berlin, Berlin, Germany.
Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2009**). "Entwurf und Aufbau eines sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio [text in German]," in *Proceedings of DAGA*, DEGA, Oldenburg, Germany, pp. 717–718.
Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2011**). "SOFiA sound field analysis toolbox," in *Proceedings of the International Conference on Spatial Audio (ICSA)*.
Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).
Blauert, J., and Rabenstein, R. (**2010**). "Loudspeaker methods for surround sound," in *Proceedings of the 57th Open Seminar on Acoustics (OSA)*, Gliwice, Poland, pp. 1–4.
Bortz, J. (**2006**). *Statistik*, 6th ed. (Springer, Berlin).
Duraiswami, R., Zotkin, D. N., Li, Z., Grassi, E., Gumerov, N. A., and Davis, L. S. (**2005**). "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," in *119th Convention of the AES*, p. 6540.
Geier, M., and Spors, S. (**2010**). "Conducting psychoacoustic experiments with the SoundScape Renderer," in *9. ITG Fachtagung Sprachkommunikation*, Bochum, Germany, pp. 1–4.
Geier, M., Spors, S., and Ahrens, J. (**2008**). "The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *124th Convention of the AES*, p. 7330, the software can be downloaded from http://spatialaudio.net/ssr/ (Last viewed 3/27/2019).
Griesinger, D. (**1996**). "Spaciousness and envelopment in musical acoustics," in *101st Convention of the AES*, Los Angeles, CA, USA, p. 4401.
Hohnerlein, C., and Ahrens, J. (**2017**). "Spherical microphone array processing in Python with the sound_field_analysis-py toolbox," in *Proceedings of DAGA*, DEGA, pp. 1–4, the software can be downloaded from https://github.com/AppliedAcousticsChalmers/sound_field_analysis-py (Last viewed 3/27/2019).
Letowski, T. (**1989**). "Sound quality assessment: Cardinal concepts," in *87th Convention of the AES*, AES, New York, p. 2825.
Lindau, A. (**2009**). "The perception of system latency in dynamic binaural synthesis," in *Proceedings of NAG/DAGA*, DEGA, Rotterdam, The Netherlands, pp. 1063–1066.
Lindau, A. (**2014**). "Binaural resynthesis of acoustical environments," Ph.D. thesis, Technische Universität Berlin, Berlin, Germany.
Lindau, A., Erbes, V., Stefan Lepa, H.-J. M., Brinkmann, F., and Weinzierl, S. (**2014**). "A spatial audio quality inventory (SAQI)," Acta Acoust. Acoust. **100**, 984–994.
McKenzie, T., Murphy, D., and Kearney, G. (**2018**). "Diffuse-field equalisation of binaural ambisonic rendering," Appl. Sci. **8**(10), 1956.
Melchior, F., Thiergart, O., Galdo, G. D., de Vries, D., and Brix, S. (**2009**). "Dual radius spherical cardioid microphone arrays for binaural auralization," in *127th Convention of the AES*, p. 7855.
Meyer, J., and Elko, G. (**2002**). "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, Orlando, FL, pp. 1781–1784.
Neidhardt, A. (**2015**). "Untersuchungen zur räumlichen Genauigkeit bei der binauralen Auralisation von Kugelarraydaten," M.Sc. thesis, Graz University of Technology, Graz, Austria.
Nowak, J., Jurgeit, K.-P., and Liebetrau, J. (**2016**). "Assessment of spherical microphone array auralizations using open-profiling of quality (OPQ)," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*.
Nowak, J., and Klockgether, S. (**2017**). "Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations," J. Acoust. Soc. Am. **142**(3), 1634–1645.
Rafaely, B. (**2005**). "Analysis and design of spherical microphone arrays," IEEE Trans. Speech Audio Process. **13**(1), 135–143.
Rasumow, E., Blau, M., Doclo, S., Hansen, M., de Par, S. V., Püschel, D., and Mellert, V. (**2013**). "Least squares versus non-linear cost functions for a virtual artificial head," in *Proceedings of Meetings on Acoustics*, Vol. 19, pp. 1–10.
Stade, P., Bernschütz, B., and Rühl, M. (**2012**). "A spatial audio impulse response compilation captured at the WDR broadcast studios," in *27th Tonmeistertagung—VDT International Convention*, pp. 1–17.
Williams, E. G. (**1999**). *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London).
Zaunschirm, M., Schörkhuber, C., and Höldrich, R. (**2018**). "Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," J. Acoust. Soc. Am. **143**(6), 3616–3627.
Zotter, F. (**2009**). "Sampling strategies for acoustic holography/holophony on the sphere," in *Proceedings of NAG/DAGA*, DEGA, Rotterdam, The Netherlands, pp. 1–4.